Revisiting the population vs phoneme-inventory correlation

Steven Moran[1,2], Daniel McCloy[1], and Richard Wright[1]
[1]University of Washington, [2]Ludwig-Maximilians-Universität Munich

Speculation about the relationship between linguistic and non-linguistic structures dates back at least a century. Sapir (1912) suggested that the influence of non-linguistic factors (such as topography, climate, flora and fauna, etc) are most clearly reflected in a language's vocabulary, but Sapir also believed that they affect the phonological and grammatical systems of languages. It is clear that certain non-linguistic contexts clearly favor differential enrichment of the lexicon, evidenced by the uneven distribution of domain-specific vocabulary in relation to the importance of those domains for different linguistic communities (e.g., Nettle, 1999). However, the relationship between phonologies and extralinguistic factors like social structure or population size has been more controversial (e.g., Atkinson, 2011; Bakker, 2004; Hay & Bauer, 2007; Lupyan & Dale, 2010; Nettle, 1999; Pericliev, 2004; Trudgill, 1996, 1997, 2002, 2004; Wichmann & Holman, 2009; Wichmann, Stauffer, Schulze, & Holman, 2008).

In one recent study published in *Language*, Hay & Bauer (2007) find a correlation between the population of speech communities and the number of phonemes in those languages. However, their methodology and results are questionable due to their statistical approach and small sample of languages (only 216). Despite their own caution in interpreting these findings, the Hay & Bauer correlations have become the basis of much other research, including the widely cited *Science* article by Atkinson (2011), which claims to show a serial founder effect in which phonological systems become smaller and less complex the further they are from the inferred point of human origin in Africa. A contrasting result is reported by Donohue & Nichols (2011), who used a much larger sample with better genealogical and areal balance, but they also used a statistical model that was inappropriate for a non-independent data structure (i.e., languages nested within language families).

Our study used a sample of 961 phonological inventories from the PHOIBLE database (Moran & Wright, 2009) and genetic and speaker population data from Ethnologue (Lewis, 2009). The data are modeled using a hierarchical mixed linear models with various subsets of log(phonemes) as outcomes, log(population) as a fixed effect predictor, and genus- and family-level language classifications as random effects predictors (genus data from WALS: Dryer & Haspelmath, 2011). We show that speaker population accounts for little to none of the variation in various measures of the phonological system (e.g., number of phonemes,

consonants, vowels, obstruents, etc). After controlling for genetic relatedness of languages, we find that some correlations (e.g., sonorant inventory size vs speaker population) are not seen in our data at all, whereas others (e.g., total phoneme inventory size vs speaker population) are marginally significant but with effect sizes so small as to be uninteresting (e.g., an increase of 1.02 phonemes per order of magnitude increase in population size, see figure 1).

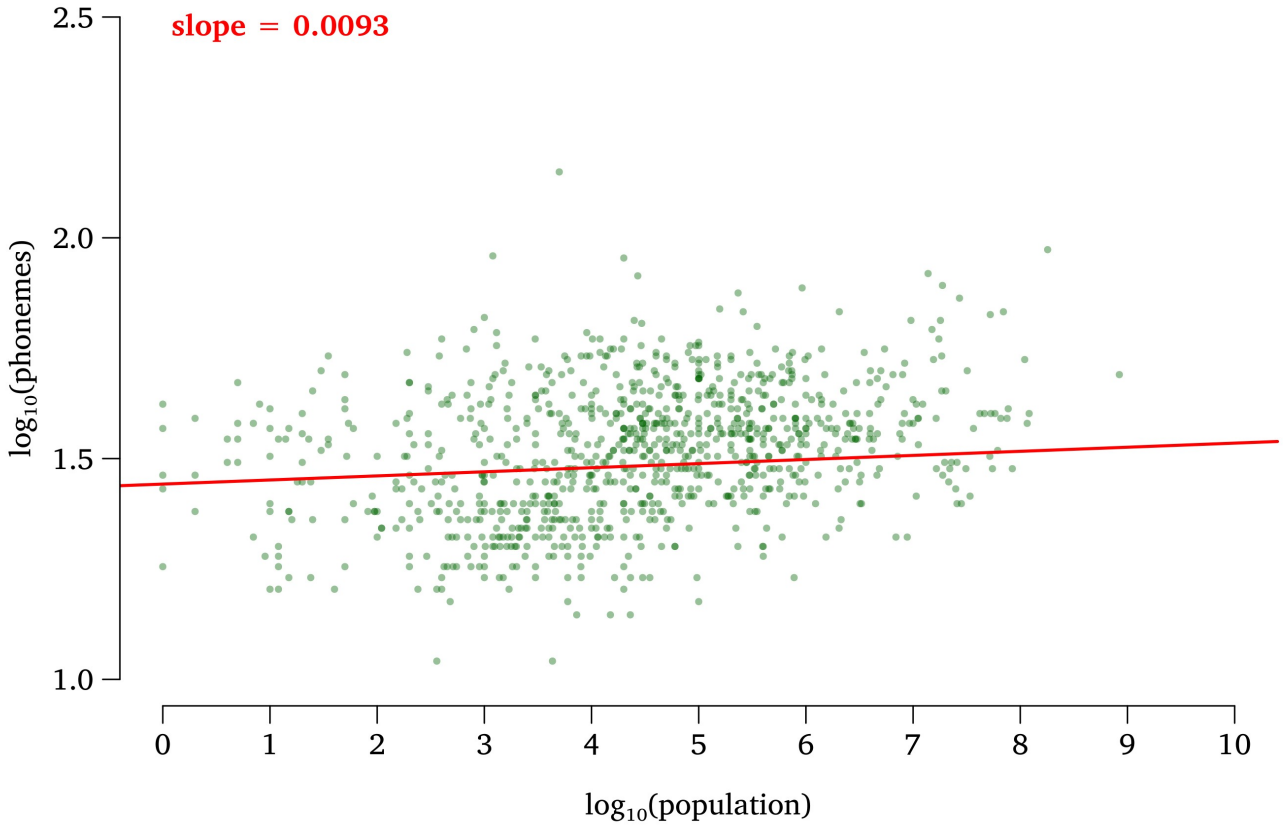

Figure 1: Overall regression of phoneme inventory size vs population size

In particular, we find the overall phonemes~population relationship to be dominated by what appear to be artefactual effects, resulting in essence from accidental facts about "outlier" languages. For example, among the small number of languages with more than a million speakers, there happen to be a few languages with high phoneme counts (e.g., Hindi, with 94 phonemes and 180 million speakers), thus pulling the regression line to exhibit a modestly positive slope. Moreover, we show that in cases where a statistically significant correlation is found, the magnitude of the predicted effect *across the entire range of the data* is still smaller than the variability seen within any one cohort of languages, when languages are grouped based on similar speaker populations (see figure 2).

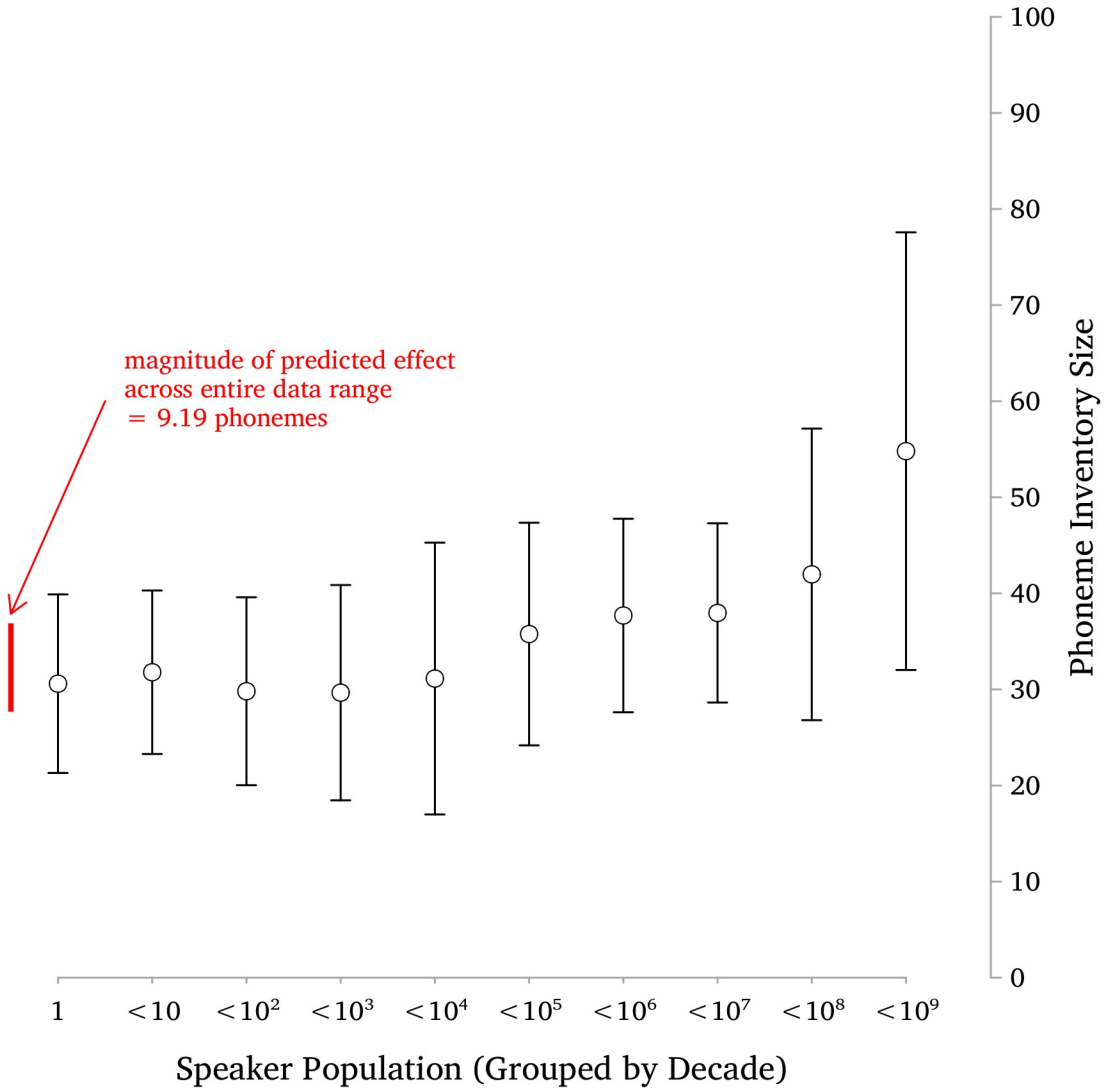

Figure 2: Means and standard deviations of phoneme inventory size by population cohort

This suggests that while the observed relationship between speaker population size and phoneme inventory size may be *statistically* significant, it is perhaps insignificant in the more usual sense of the word. Finally, we show that in cases where we see an overall positive trend across all languages, the trend is not preserved within families. Indeed, some language families are best modeled by inverse correlations between speaker population and phoneme inventory size, whereas regressions within other families show no correlation whatsoever (see figure 3). This finding supports the view that the correlations we do see at the overall level are indeed artefactual.

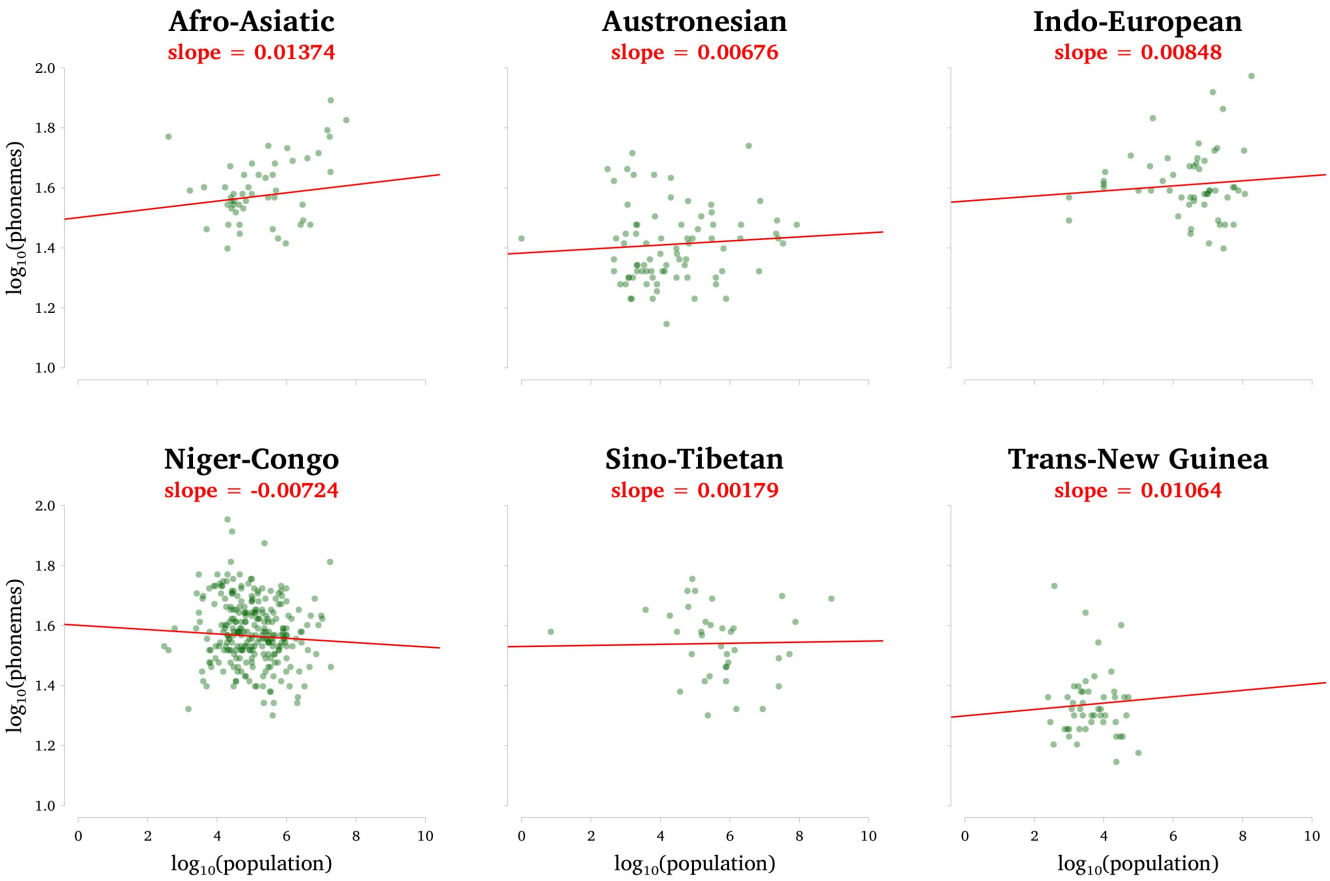

Figure 3: Within-family regressions for the six largest language families

Our results are discussed in light of the previous literature. We argue that the spurious correlations in other studies are due to (1) failing to control for genetic relatedness of languages, (2) samples skewed by small size or over-representation of certain language families, (3) case-based reasoning that lacks statistical rigor, or (4) some combination of the above. We also raise methodological questions about model interpretation and hypothesis testing: specifically, we reason that although it is possible that factors effecting population (e.g., immigration, cultural assimilation, war, disease) might lead to phonological change, it is by no means obvious that phonological change is a necessary consequence of population change. Following a series of arguments by Trudgill (summarized in Trudgill, 2011), we reason that a number of interrelated linguistic and sociolinguistic factors may in fact be relevant (e.g., differences between the phonologies of the languages in question, relative population sizes, details of the language contact situation, etc), and therefore neither would we expect phonological change to result from population change alone, nor would we expect population size to consistently index the myriad interacting factors that are likely at play.

References

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, *332*(6027), 346–349. doi:10.1126/science.1199295

Bakker, P. (2004). Phoneme inventories, language contact, and grammatical complexity: A critique of Trudgill. *Linguistic Typology*, *8*(3), 368–375. doi:10.1515/lity.2004.8.3.368

Donohue, M., & Nichols, J. (2011). Does phoneme inventory size correlate with population size? *Linguistic Typology*, *15*, 161–170. doi:10.1515/LITY.2011.011

Dryer, M. S., & Haspelmath, M. (Eds.). (2011). *The world atlas of language structures online*. Munich: Max Planck Digital Library. Retrieved from http://wals.info/

Hay, J., & Bauer, L. (2007). Phoneme inventory size and population size. *Language*, *83*(2), 388–400.

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (16th ed.). Dallas, TX: SIL International. Retrieved from http://www.ethnologue.com/

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*(1), e8559. doi:10.1371/journal.pone.0008559

Moran, S., & Wright, R. (2009). *Phonetics information base and lexicon (PHOIBLE)*. Seattle, WA. Retrieved from http://phoible.org/

Nettle, D. (1999). *Linguistic diversity*. Oxford: Oxford University Press.

Pericliev, V. (2004). There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language. *Linguistic Typology*, *8*(3), 376–383. doi:10.1515/lity.2004.8.3.376

Sapir, E. (1912). Language and environment. *American Anthropologist*, New Series, *14*(2), 226–242.

Trudgill, P. (1996). Dialect typology: Isolation, social network and phonological structure. In G. R. Guy, C. Feagin, D. Schiffrin, & J. Baugh (Eds.), *Towards a social science of language: Papers in honor of William Labov* (Vols. 1-2, Vol. 1, pp. 3–21). Amsterdam: John Benjamins.

Trudgill, P. (1997). Typology and sociolinguistics: Linguistic structure, social structure and explanatory comparative dialectology. *Folia Linguistica*, *31*(3/4), 349–360. doi:10.1515/flin.1997.31.3-4.349

Trudgill, P. (2002). Linguistic and social typology. *The handbook of language variation and change* (pp. 707–728). Oxford: Blackwell.

Trudgill, P. (2004). Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology*, *8*(3), 305–320. doi:10.1515/lity.2004.8.3.305

Trudgill, P. (2011). Social structure and phoneme inventories. *Linguistic Typology*, *15*, 155–160. doi:10.1515/LITY.2011.010

Wichmann, S., & Holman, E. W. (2009). Population size and rates of language change. *Human Biology*, *81*(2/3), 259–274. doi:10.1353/hub.0.0059

Wichmann, S., Stauffer, D., Schulze, C., & Holman, E. W. (2008). Do language change rates depend on population size? *Advances in Complex Systems*, *11*(3), 357. doi:10.1142/S0219525908001684